Enriched Lexical Ontologies: 
Adding new knowledge and new scope to old linguistic resources

Tony Veale 
School of Computer Science, 
UCD (NUI-D) 
Tony.Veale@UCD.ie
Overview: The WHY and WHAT of Enriched Lexical Ontologies?

What is a Lexical Ontology

An ontology of lexical(-ized) concepts, used in NLP, serving as a lexical semantics

Coverage dictated by lexical concerns/domain vocabulary

E.g., Princeton WordNet aims to capture the lexical breadth of English

Broad coverage (usually) dictates “shallow” treatment

E.g., PWN not axiomatized (unlike Cyc), but see OntoClean treatment of PWN

Can be driven by specific theory of lexical semantics

E.g., Generative Lexicon (GL) theory can be used to organize a lexical ontology

What and Why?

Term Coverage

Lightweight

Theoretical basis
Overview: The (intended) structure of this course

**Introduction to Enriched Lexical Ontologies**
Overview of Lexical Ontologies, approaches to and sources of ont. enrichment

**Lexical Acquisition from the Web / Wikipedia**
A dynamic lexicon that enriches itself from Wikipedia; “portmanteau” words

**A Relational Overlay for WordNet**
Overlaying a framework of relations onto senses in PWN to support inference

**Making WordNet Functional and Content-Sensitive**
Enriching PWN with “category functions”; acquiring diagnostic features from Web
Some Well-known Lexical Ontologies

• Princeton WordNet  *(Miller '90, Fellbaum '98)*
  [wordnet.princeton.edu](http://wordnet.princeton.edu)
  Lightweight ontology of nouns, verbs and adjectives; good noun taxonomy

• Euro-WordNet(s)  *(Piek Vossen)*
  [www.illc.uva.nl/EuroWordNet](http://www.illc.uva.nl/EuroWordNet)
  PWN-style WordNets for Euro-languages (even English!), additional rel-types

• CYC  *(Lenat & Guhu and countless "cyclists")*
  [www.cycorp.com](http://www.cycorp.com)
  An ambitious large-scale ontology of common-sense and linguistic knowledge

• SUMO  *(Pease and Niles)*
  [www.ontologyportal.org](http://www.ontologyportal.org)
  Suggested Upper-Merged Ontology (KIF; linked to PWN; an upper-ontology)
### Some Other Lexical Ontologies

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DOLCE</strong></td>
<td>(Gaurino, Gangemi) Descriptive Ontology for Linguistic &amp; Cognitive Engineering; formally engineered</td>
<td>dolce.semanticweb.org</td>
</tr>
<tr>
<td><strong>LDOCE</strong></td>
<td>(Longman’s Dict. of Contemp. English) Machine-readable dictionary, with semantic “box codes”</td>
<td><a href="http://www.ldoce.com">www.ldoce.com</a></td>
</tr>
<tr>
<td><strong>Mikrokosmos</strong></td>
<td>(Nirenburg, Raskin ’87) Text-Meaning Representation for Machine Translation systems</td>
<td>crl.nmsu.edu/Research/Projects/mikro</td>
</tr>
</tbody>
</table>
Princeton WordNet: a network of word-senses as synonym-sets

Concepts are SynSets (Synonym Sets)  Nouns hierarchy is organized by isa links
Taxonomic Structure in HowNet is more impoverished than in PWN

The HowNet category human | 人 has several thousand direct hyponyms!
However, HowNet uses Conceptual Graph Definitions that can be mapped

If only more definitions in HowNet were so amenable to structure-mapping

\[
\text{Struct Hash: } \{?:{\text{ill|病态 } :OfPart=\{?\},experiencer=\{\sim\},scope=\{?\}}\}
\]
HowNet: Property / Modifier Taxonomy

- Appearance Value
- Sharpness Value
- Taste Value
- Posture Value

E.g., (samurai)武士 ⇒ Samurai:Courage Value = valiant 武
Exploiting Semantic Transparency in other Languages/Ontologies

Espresso = 

濃咖啡

Deep (isa HueValue)
Strong (isa IntensityValue)
Rich (isa TasteValue)
Concentrated (isa ConcentrationValue)
Thick (isa DensityValue)

coffee (isa drinks)
Enrichment as the Construction of Slot-filler Frame-Structures

WordNet headword
as frame-name

HowNet parent of orthographic modifier
yields a Slot-Name

Translation of HowNet modifier
yields a Slot-Filler

CAUTION:
This enrichment relies on the sensibility of the HowNet ontology
+ transparency of orthographic modifiers
### Enriched (third-party) Versions of WordNet

<table>
<thead>
<tr>
<th>Version</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WordNet Domains</strong> (Magnini &amp; Cavaglià '00)</td>
<td>~200 domain labels (Sport, etc.) to complement PWN's own domain markings</td>
<td>wndomains.itc.com</td>
</tr>
<tr>
<td><strong>WordNet-Affect</strong> (Strapparava &amp; Valitutti, '04)</td>
<td>Additional labels for affective concepts (moods, emotions, traits, cog. states)</td>
<td>wndomains.itc.com</td>
</tr>
<tr>
<td><strong>SentiWordNet</strong> (Esuli &amp; Sebastiani, '06)</td>
<td>Assigns sentiment scores (for positivity, negativity, objectivity) to PWN synsets</td>
<td>sentiwordnet.isti.cnr.it</td>
</tr>
<tr>
<td><strong>eXtended WordNet</strong> (Moldovan &amp; Harabagiu)</td>
<td>Parsed logical forms and sense-tags for sentential content of PWN glosses</td>
<td>xwn.hlt.utdallas.edu</td>
</tr>
</tbody>
</table>
Categorization in WordNet: How Well is it Supported?

Categorization requires each sense be given a “radial” category-membership “function”
(Some obvious) Problems with WordNet

- Small set of non-hierarchical relations
  part/whole, etc.
  Not sufficient for complex reasoning (analogy, common-sense inference, etc.)

- No distinction between polysemy and homonymy
  Co-active senses
  No links connecting related senses of the same word (metonymy, reg. polysemy)

- No explicit marking of metaphoric senses
  Figurative Language
  E.g., no marking of “bread” (qua money) as a metaphoric sense extension

- Glosses are for human, not machine, consumption
  Implicit meaning
  Not in logical form; not sense-tagged; not consistent across definitions
Sense-Discrimination promotes Meaning Disintegration

MONOPOLY

(MARKET + DOMINANCE)

"(ECONOMICS) A market in which there are many buyers but only one seller"

"EXCLUSIVE CONTROL OR POSSESSION OF SOMETHING"
Polysemy and Homonymy are not marked in WordNet

- **OBJECT**
  - isa
  - **ARTIFACT**
  - isa
  - **FACILITY**
  - isa
  - **DEPOSITORY**
  - isa
  - **BANK (8)**
    - “A BUILDING IN WHICH COMMERCIAL BANKING IS TRANSACTED”

- **LOCATION**
  - isa
  - **GROUP**
    - isa
  - **ORGANIZATION**
    - isa
  - **INSTITUTION**
    - isa
  - **BANK (10)**
    - “A FINANCIAL INSTITUTIONS THAT ACCEPTS DEPOSITS …”

- **OBJECT**
  - isa
  - **GEOLOGICAL FORMATION**
    - isa
  - **SLOPE**
    - isa
  - **BANK (13)**
    - “SLOPING LAND”
## Some (more) Problems with WordNet

<table>
<thead>
<tr>
<th>Problem</th>
<th>Addressed in this course</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets are NOT Conceptual Categories</td>
<td>Addressed in this course</td>
<td>No membership criteria; no intensional definitions</td>
</tr>
<tr>
<td>Not Cognitively Structured</td>
<td>Addressed in this course</td>
<td>No category prototypes, no radial categories, no family resemblances</td>
</tr>
<tr>
<td>Not Contextualized</td>
<td>Addressed in this course</td>
<td>Categories stretch and adapt to changing language contexts; but not in PWN</td>
</tr>
<tr>
<td>No Compartmentalization of meanings</td>
<td>Important, but not here</td>
<td>E.g., “Sherlock Holmes” is both fictional content and a detective (qua person)</td>
</tr>
</tbody>
</table>
Conceptual Profiling: What is Base and What is Profiled?

How to find cognitively profiled features? How to exploit them via enrichment?

- Example: Espresso

- Features:
  - Black
  - Gourmet
  - Trendy
  - Steamy
  - Cold
  - Satisfying
  - Concentrated
  - Creamy
  - Rich
  - Small
  - Intense
  - Strong
  - Italian
  - Expensive
Consider \{Skilled\ _Worker\} in Princeton WordNet

What are the membership criteria for being a “Skilled Worker”?

*PWN doesn’t say …*

Does any skill suffice? Must it be job-related?

What about actors? Aren’t they skilled?

In some contexts (e.g., for green cards), Actors and Concern Pianists are “skilled” workers
### Sources of Enrichment for WordNet

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WordNet itself</strong></td>
<td>WordNet-internal&lt;br&gt;Glosses (as in eXtended WordNet), Synsets (Aprejean’s regular polysemy), etc.</td>
</tr>
<tr>
<td><strong>Other Ontologies</strong></td>
<td>WordNet-external&lt;br&gt;E.g., HowNet, Cyc, Microsoft’s MindNet, Euro-WordNets, Domain Ontologies …</td>
</tr>
<tr>
<td><strong>Plain Text Corpora</strong></td>
<td>WordNet-external&lt;br&gt;Information-extraction techniques to acquire lexemes, meanings from texts</td>
</tr>
<tr>
<td><strong>Semi-Structured Texts and Databases</strong></td>
<td>WordNet-external&lt;br&gt;E.g., Wikipedia (links + categories + templates), LDOCE and other MRDs, …</td>
</tr>
</tbody>
</table>
E.g., WordNet-Internal: Sense Linkage Via Bridging Compounds

- **Olive (1)**: “Small ovoid fruit of the European olive tree”
- **Olive (2)**: “Evergreen tree cultivated in the Mediterranean since…”
- **Olive (3)**: “Hard yellow often variegated wood of …”
WordNet-Internal: Sense Linkage Via Gloss-Referencing

“A stick of black carbon material used for DRAWING”

“Drawing made with charcoal”

Extended WordNet sense-tagged glosses make this process more reliable
## WordNet-Internal: Linking Senses of Polysemous words

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Coverage</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridging Compounds</td>
<td>15%</td>
<td>95%</td>
</tr>
<tr>
<td>Gloss Cross-Referencing</td>
<td>76%</td>
<td>89%</td>
</tr>
<tr>
<td>Hierarchical Reinforcement</td>
<td>11%</td>
<td>76%</td>
</tr>
<tr>
<td>Blend Recruitment</td>
<td>3%</td>
<td>89%</td>
</tr>
<tr>
<td>Morphosemantic Linking</td>
<td>3%</td>
<td>93%</td>
</tr>
<tr>
<td>All Strategies (25)</td>
<td>98%</td>
<td>85%</td>
</tr>
</tbody>
</table>

* compared to hand-coded exceptions-list (PWN “cousins list”)

---

## Percentage and Meanings

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Coverage</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridging Compounds</td>
<td>15%</td>
<td>95%</td>
</tr>
<tr>
<td>Gloss Cross-Referencing</td>
<td>76%</td>
<td>89%</td>
</tr>
<tr>
<td>Hierarchical Reinforcement</td>
<td>11%</td>
<td>76%</td>
</tr>
<tr>
<td>Blend Recruitment</td>
<td>3%</td>
<td>89%</td>
</tr>
<tr>
<td>Morphosemantic Linking</td>
<td>3%</td>
<td>93%</td>
</tr>
<tr>
<td>All Strategies (25)</td>
<td>98%</td>
<td>85%</td>
</tr>
</tbody>
</table>

* compared to hand-coded exceptions-list (PWN “cousins list”)
Ontologies like PWN are Semi-Structured: Some meaning in flat text

WordNet-Internal:
Taxonomy can be improved by extraction from glosses

E.g., Alpha = “The 1st letter of the Greek Alphabet”
Using PWN: An Explicit Taxonomy with Implicit Gloss Content

WordNet glosses can be shallow parsed to reveal pivotal categorization features
WordNet-Internal: Making Latent Gloss Content Explicit

\{\text{Alpha}\} \equiv \text{“The 1\textsuperscript{st} letter of the \underline{Greek} Alphabet”}

WordNet-External: We look at a corpus-based approach to feature-acquisition in a later class.
<table>
<thead>
<tr>
<th></th>
<th>Deity to Deity Mapping Task</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td></td>
</tr>
<tr>
<td>Static WN</td>
<td>0.115</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>representations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic WN</td>
<td>0.935</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>representation</td>
<td>(+ gloss-feature reification)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                  | Letter to Letter Mapping Task |            |          |
|                  | Precision                   | Recall     |
| Static WN        | 0.04                        | 0.98       |
| representations  |                             |            |
| Dynamic WN       | 0.96                        | 0.98       |
| representation   | (+ gloss-feature reification)|            |

E.g., Greek to Roman gods, Hindu to Semitic gods, etc.

Greek to Hebrew letters, and Hebrew to Greek letters.
WordNet-External **Enrichment Schemes**

- **Corpus acquisition of hypernyms & meronyms**
  
  Seek reliable text patterns: “antidepressants like Prozac, Zoloft and XYZ …”

- **Corpus/Web acquisition of GL Qualia structures**

  E.g., Telic of iPod: “used to download/store/play music …”

- **Semi-structured Wikipedia content (categories)**

  Enriches PWN with Wikipedia place-names, proper entities [*also: templates*]

- **Acquire diagnostic category features from Web**

  Build context-sensitive category functions from salient features of concepts

---

Exploits common constructions and “patterns of coining” (Kay)
To Recap: The (intended) structure of this course

<table>
<thead>
<tr>
<th>Introduction to Enriched Lexical Ontologies</th>
<th>Done (?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview of Lexical Ontologies, approaches to and sources of ont. enrichment</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexical Acquisition from the Web / Wikipedia</th>
<th>Tomorrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>A dynamic lexicon that enriches itself from Wikipedia; “portmanteau” words</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A Relational Overlay for WordNet</th>
<th>Day #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlaying a framework of relations onto senses in PWN to support inference</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Making WordNet Functional and Content-Sensitive</th>
<th>Day #4/ #5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriching PWN with “category functions”; acquiring diagnostic features from Web</td>
<td></td>
</tr>
</tbody>
</table>
Cultural Learnings of Wikipedia
For make benefit glorious resource of WordNet

Tony Veale
School of Computer Science,
UCD (NUI-D)
Tony.Veale@UCD.ie
## Neologisms and the Lexicon

<table>
<thead>
<tr>
<th>Neologisms are novel word forms (New-Word)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., “soccer mom”, “neocon”, “gastropub”, “affluenza”, “chicken hawk”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neologisms enter the lexicon gradually</th>
<th>Bubbling Under</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can be used without gaining widespread recognition for years</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wikipedia</th>
<th>Reflects Cultural Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Updated more frequently than print dictionaries, captures the “Zeitgeist”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ZeitGeist</th>
<th>Neologism Harvester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trawls Wikipedia looking for neologisms (w.r.t WordNet) and their meanings</td>
<td></td>
</tr>
</tbody>
</table>
### Lexicographic Approaches to Neologism Analysis

<table>
<thead>
<tr>
<th>Approach</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory Lexicography</td>
<td>Post-Hoc Analysis</td>
<td>“Affluenza&quot; = “Affluent&quot; + “Influenza&quot; (overlap: “…fluen…”)</td>
</tr>
<tr>
<td>Predictive Lexicography</td>
<td>Pre-Hoc Creativity</td>
<td>“Chrono-“ (= time) + “-onaut“ (=traveller) = “<em>chrononaut</em>” (a “time traveller”)</td>
</tr>
</tbody>
</table>

*Explanatory Lexicography at Work* — Seeks to dissect and explain neologisms after they have been identified in text.

*Predictive Lexicography at Work* — Seeks to predict and verify neologisms using principles of word-formation.
<table>
<thead>
<tr>
<th>Type</th>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni-Directional Links</td>
<td>A→B</td>
<td>Headwords accompanied by text article that links to relevant other articles</td>
</tr>
<tr>
<td>Bidirectional Reciprocated</td>
<td>A←B</td>
<td>Two headwords cross-reference each other, indicating strong mutual association</td>
</tr>
<tr>
<td>Links</td>
<td>A→B ; C</td>
<td>The Wikipedia article for A contains a link to B followed directly by a link to C</td>
</tr>
<tr>
<td>Text-Free Analysis (No Parsing</td>
<td>Topological Approach</td>
<td>The text of each article is ignored; only topology of headword xrefs is used</td>
</tr>
</tbody>
</table>
# Compound Terms

- **Simplest: ADJ-Noun and Noun-Noun collocations**
  - Prevalent in English/ PWN
  - Yokes two branches of an ontology together: e.g., Religious-Music, Chinese-Cuisine

- **General Structure**
  - **Endocentricity**
  - \(<\text{Modifier : Head}>\), where Head denotes a hypernym of the compound meaning

- **Awkward Exceptions**
  - **Exocentricity**
  - \(<\text{Modifier : Head}>\) does not denote a hyponym of \{Head\}, e.g., “hammer head”

- **Inter-Compound Relationships**
  - **Specialization of Parts**
  - Applied-Science :: Applied-Physics but Religious-Music :: Christian-Music
Compound Schema I: Head Specialization

Endocentric form: $\alpha_\beta$

$\alpha_\beta \rightarrow \alpha_\gamma \land \beta \text{ isa } \gamma$

$\alpha_\beta \text{ isa } \alpha_\gamma$

PWN senses of $\beta$ and $\gamma$ given by this relationship

Examples

- Fantasy_Sport $\rightarrow$ “Fantasy_Football and Football isa Sport (in PWN)

- Veterinary_Science $\rightarrow$ Veterinary_Medicine and Medicine isa Science (in PWN)
Compound Schema II: Head Anchoring

\[ \alpha_\beta \rightarrow \gamma \land \beta \text{ isa } \gamma \]

Anchored in non-compound PWN senses of \( \beta \) and \( \gamma \) given by this relationship

- Examples
  - Applied_Statistics \( \rightarrow \) Science and Statistics isa Science (in PWN)
  - British_Rail \( \rightarrow \) Railway and Rail isa (shorthand for) Railway (in PWN)
Compound Schema III: Compound Expansion

\[ \alpha_\beta \rightarrow \alpha \land \alpha \text{ isa } \beta \]

\[ \alpha_\beta \text{ synonym of } \alpha \]

Examples

- Oscar_award → Oscar and Oscar isa award (in PWN)
- Semitic_language → Semitic and Semitic isa Language (in PWN)
Compound Schema IV: Compound Conflation

\[ \alpha_\beta \rightarrow \beta_\gamma \land \beta \text{ isa } \gamma \]

\[ \alpha_\beta \text{ isa } \beta_\gamma \]

**Examples**

- Touch_rugby $\rightarrow$ Rugby_football and Rugby isa football (in PWN)
- Pop_punk $\rightarrow$ Punk_Rock and Punk isa Rock (in PWN)

\( \beta_\gamma \) may be a new compound, previously learned.
Compound Schema V: Compound Opposition

\[
\alpha_\beta \rightarrow \alpha_\gamma \land \gamma \text{ antonym } \beta
\]

\[
\alpha_\beta \text{ antonym of } \alpha_\gamma
\]

Examples:
- Second_language → First_language and Second antonym of First (in PWN)
- Synthetic_geometry → Analytic_geometry and Synthetic opposes Analytic
Compound Schema VI: Head Expansion

\[ \alpha_\beta \rightarrow \gamma_\beta \land \gamma_\beta \text{ synonym of } \beta \Leftarrow \]

\[ \alpha_\beta \text{ isa } \gamma_\beta \Leftarrow \]

Examples

- Escort_Carrier $\rightarrow$ Aircraft_carrier and Aircraft_carrier syn. of Carrier
- Simple_majority $\rightarrow$ Absolute_Majority and Absolute_Majority syn. Majority
Compound Schema VII: Modifier Specialization

\[ \alpha_\beta \rightarrow \gamma_\beta \land \alpha \text{ mod-isa } \gamma \]

\[ \alpha_\beta \text{ isa } \gamma_\beta \]

Examples

- Truck_racing $\rightarrow$ Auto_racing and Truck is a \textbf{automotive Vehicle} (in PWN)
- Hindu_music $\rightarrow$ Religious_Music and Hindu is a \textbf{Religious person} in (PWN)

$+$ vice versa

PWN senses of $\alpha$ and $\gamma$ given by this relationship
**Compound Schema VIII : Catch-all Coordination**

\[
\alpha_\beta \rightarrow \gamma_\beta \land \gamma_\beta \text{ isa } \beta \\
\alpha_\beta \text{ coordinate of } \gamma_\beta
\]

- **Examples**
  - Financial_mathematics \rightarrow applied_mathematics (isa mathematics)
  - Constructed_language \rightarrow natural_language (isa language)

**Weakest schema**

**PWN senses of \(\beta\) and \(\gamma_\beta\) given by this relationship**
# Endocentric Compound Evaluation: Results

For 10,899 Wikipedia headwords matching one or more compound schema:

<table>
<thead>
<tr>
<th>Schema</th>
<th># Headwords</th>
<th># Error</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>14%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>II</td>
<td>12%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>III</td>
<td>13%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>IV</td>
<td>4%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>V</td>
<td>1%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>VI</td>
<td>3%</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>VII</td>
<td>15%</td>
<td>7%</td>
<td>.93</td>
</tr>
<tr>
<td>VIII</td>
<td>70%</td>
<td>0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

*E.g.,* Dutch_language isa German_language

*E.g.,* Supreme_Court isa state_court
### In-depth Analysis: Portmanteau Words

<table>
<thead>
<tr>
<th>Portmanteau (double-pocket) words</th>
<th>Lewis Carroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Textual blend of two different words, e.g., “Bollywood”, “Infomercial”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General Structure</th>
<th>Prefix + Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>One words contributes a prefix, the other a suffix (but, e.g., “Modem”)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True Portmanteaux</th>
<th>Double-Scope Blends</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither component word is present in its entirety, (e.g., “metrosexual”)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Impure Portmanteau</th>
<th>Broad Exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., “Wikipedia”, (but not “Wiktionary”), “Gastropub”, “Feminazi”, ...</td>
<td></td>
</tr>
</tbody>
</table>
## Taxonomic Connectives

<table>
<thead>
<tr>
<th>Precise Taxonomic Placement</th>
<th>Pure ISA</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., “Superhero” = “Super-“ + “Hero“ ⇒ Superhero ISA Hero</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approximate Taxonomic Placement</th>
<th>Hedging</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., “Spintronics” = “Spin“ + “Electronics“ ⇒ Spintronics hedges Electronics</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disambiguation</th>
<th>Sense Priming</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., Which sense of “hero“ does “Superhero“ extend? (not a sandwich obviously)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hedging Supports Figurative Portmanteaux</th>
<th>Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Affluenza“ = “Affluence“ + “Influenza“ but Affluenza NOT-ISA Influenza</td>
<td></td>
</tr>
</tbody>
</table>
General Approach: **Two-Pass Harvesting**

A Textual, String-Matching Approach

Let $\alpha\beta$ represent the general form of a headword; analyse with schemata

The Topological Context

The set of Wikipedia cross-references for a given headword $\alpha\beta$

Pass I: Learning from Easy Cases

Harvest obvious examples (in rich contexts) first, and learn from these cases

Pass II: Applying learnt patterns to Hard Cases

When topological context is insufficient, use experience of easy cases as a guide
Portmanteau Schema I: Explicit Extension

\[ \alpha \beta \rightarrow \beta \land \alpha \beta \rightarrow \alpha \gamma \]

\[ \alpha \beta \text{ isa } \beta \]

- Examples
  - Gastropub ("Gastropub" $\rightarrow$ "pub" and "Gastropub" $\rightarrow$ "Gastronomy")
  - Feminazi ("Feminazi" $\rightarrow$ "Nazi" and "Feminazi" $\rightarrow$ "Feminism")
Portmanteau Schema II: Suffix Alternation

\[ \alpha \beta \rightarrow \alpha \gamma \land \beta \leftrightarrow \gamma \]

\[ \alpha \beta \text{ hedges } \alpha \gamma \]

- Examples
  - “man” ↔ “boy” “woman” ↔ “girl” “bit” ↔ “byte” “toxin” ↔ “bacteria”
  - Fangirl (“Fangirl” → “Fanboy” and “boy” → “girl”)

Easy
Portmanteau Schema III: Partial Suffix

\[
\alpha \beta \rightarrow \gamma \beta \land (\alpha \beta \rightarrow \alpha \lor \alpha \beta \rightarrow \delta \rightarrow \alpha)
\]

\[
\alpha \beta \text{ hedges } \gamma \beta
\]

- Examples
  - "Metrosexual" $\rightarrow$ "Heterosexual" $\land$ "Metrosexual" $\rightarrow$ "Metro"
  - "Pomosexual" $\rightarrow$ "Homosexual" $\land$ "Pomosexual" $\rightarrow$ "Postmodernism" $\rightarrow$ "pomo"
Portmanteau Schema IV: Consecutive Blending

\[ \alpha\beta \rightarrow \alpha\gamma ; \delta\beta \]

\[ \alpha\beta \text{ hedges } \delta\beta \]

- Examples
  - “sharpedo” \(\rightarrow\) “shark” ; “torpedo” hedges “torpedo”
  - “Spanglish” \(\rightarrow\) “Spanish” ; “English” hedges “English”
Portmanteau Schema IVa: Partial Suffix

\[ \alpha \beta \rightarrow \alpha \gamma ; \delta \beta \land \alpha \beta \rightarrow \text{portmanteau} \]

\[ \alpha \beta \text{ hedges } \gamma \beta \]

- Examples

“Spork” \( \rightarrow \) “Spoon” \( \land \) “Fork” \( \text{ hedges } \) “English” \( \text{ (2 characters)} \)

“Sporger” \( \rightarrow \) “Spam” \( \land \) “Forgery” \( \text{ hedges } \) “Forgery” \( \text{ (2 characters)} \)
Portmanteau Schema V: Suffix Completion

\[ \alpha \beta \rightarrow \gamma \beta \land \gamma \beta \in \mathcal{H}_{III} \land \beta \in \mathcal{S}_{III} \]

\[ \alpha \beta \text{ hedges } \gamma \beta \]

- Examples
  - "Retrosexual" \( \rightarrow \) "Metrosexual" \( \land \) "sexual" \( \in \mathcal{S}_{III} \)
  - "crippleware" \( \rightarrow \) "malware" \( \land \) "ware" \( \in \mathcal{S}_{III} \)

\( \mathcal{H}_{III} = \text{headwords} \)
\( \mathcal{P}_{III} = \text{prefixes} \)
\( \mathcal{S}_{III} = \text{suffixes} \)
Portmanteau Schema VI: Separable Suffix

\[ \alpha \beta \rightarrow \beta \wedge \alpha \in (\mathcal{P}_I \cup \mathcal{P}_{II} \cup \mathcal{P}_{III}) \]

\[ \alpha \beta \text{ is a } \beta \]

• Examples

“Gastroshop” $\rightarrow$ “shop” $\wedge$ “gastro” $\in \mathcal{P}_{III}$

“antiprism” $\rightarrow$ “prism” $\wedge$ “anti” $\in \mathcal{P}_{III}$
Portmanteau Schema VII: Prefix Completion

\[
\alpha \gamma \rightarrow \alpha \land <\gamma, \beta> \in \mathcal{T}_I
\]

\[
\alpha \beta \text{ isa } \beta
\]

- **Examples**

  "Logicism" $\rightarrow$ "logic" $\land$ <"ism", "Nazi"> $\in \mathcal{T}_I$ so Logicnazi ISA Nazi

  "Psychology" $\rightarrow$ "psycho" $\land$ <"ology", "technology"> $\in \mathcal{T}_I$
Portmanteau Schema VIII: Recombination

\[ \alpha \beta \rightarrow \alpha \gamma \land \alpha \beta \rightarrow \delta \beta \land \alpha \in \mathcal{P}_{III} \land \beta \in \mathcal{S}_{III} \]

\[ \alpha \beta \text{ hedges } \delta \beta \]

- **Examples**
  
  - "geonym" \(\rightarrow\) "geography" \(\land\) "geonym" \(\rightarrow\) "toponym" (hedges geonym)
  
  - "dubtitle" \(\rightarrow\) "dubbed" \(\land\) "dubtitle" \(\rightarrow\) "subtitle" (hedges subtitle)
### Evaluation: Set-Up

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia contains 152,060 atomic headwords at this time of download</td>
<td></td>
</tr>
<tr>
<td>WordNet Version: 1.6</td>
<td>WordNet</td>
</tr>
<tr>
<td>Few differences were achieved using WordNet 2.1 instead</td>
<td></td>
</tr>
<tr>
<td># of Wikipedia Entries Matching a ZeitGeist schema</td>
<td>Initial Selection</td>
</tr>
<tr>
<td>4676 headwords match at least one Zeitgeist schema</td>
<td></td>
</tr>
<tr>
<td>Metaphor is a useful tool for ontological development</td>
<td>Pre-Filtering</td>
</tr>
<tr>
<td>1385 already in WordNet; 1083 analyses yield non-PWN parent or hedge</td>
<td></td>
</tr>
</tbody>
</table>
Portmanteau Evaluation: Results

For 2048 Wikipedia headwords matching one or more Portmanteau schema:

<table>
<thead>
<tr>
<th>Schema</th>
<th># Headwords</th>
<th># Error</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>710</td>
<td>11</td>
<td>.985</td>
</tr>
<tr>
<td>II</td>
<td>144</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>III</td>
<td>330</td>
<td>5</td>
<td>.985</td>
</tr>
<tr>
<td>IV</td>
<td>82</td>
<td>2</td>
<td>.975</td>
</tr>
<tr>
<td>V</td>
<td>161</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>VI</td>
<td>321</td>
<td>16</td>
<td>.95</td>
</tr>
<tr>
<td>VII</td>
<td>340</td>
<td>32</td>
<td>.90</td>
</tr>
<tr>
<td>VIII</td>
<td>320</td>
<td>11</td>
<td>.965</td>
</tr>
</tbody>
</table>

E.g.,
Rubbergate from Watergate

E.g.,
Retrosexual from Metrosexual
## Conclusions

- **A Linguistics-Lite Approach to Neologisms is Feasible**
  - Lightweight
  - No text parsing or morphological analysis; all relevant morphemes are learned

- **Taxonomic Hedging** is required
  - Uncertainty
  - Word-forms are not deterministic w.r.t. taxonomic placement, approx. needed

- **Link Topology offers context-specific insights**
  - Grounding
  - E.g., Microsurgery $\rightarrow$ Microscopy + surgery $\Rightarrow$ "surgery done with a microscope"

- **Not biased towards English**
  - Multilingual
  - Linguistics-lite means no language-bias - applicable to other languages / wikis?
Palimpsest: Overlaying Multiple Knowledge-sources onto PWN for an Enriched, Flexible WordNet

Tony Veale
School of Computer Science,
UCD (NUI-D)
Tony.Veale@UCD.ie
Why Enrichment? Desirable Features of a Robust **Lexical Ontology**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comprehensive Scale</strong></td>
<td>Size</td>
</tr>
<tr>
<td><strong>A Rich, Hierarchically-Organized Category Structure</strong></td>
<td>Organization</td>
</tr>
<tr>
<td><strong>A Rich Relational Structure</strong></td>
<td>Connectivity</td>
</tr>
<tr>
<td><strong>Graded Levels of Salience</strong></td>
<td>Texture</td>
</tr>
</tbody>
</table>
# Combining Knowledge Resources to Enrich Lexical Ontologies

- **Princeton WordNet**
  - ISA Hierarchy, glosses
  - Good taxonomy for noun-concepts; text glosses can be mined for extra content

- **Corpora and The text of the WWW**
  - *discussed tomorrow*
  - Diagnostic features (e.g., *surgeons are precise*) extracted from the WWW

- **Wikipedia**
  - Salience and Topicality
  - A rich topology of cross-reference linkages between articles (headwords)

- **Palimpsest (a hand-crafted KR backbone)**
  - Relational Triples
  - An explicit relational framework (of two-place relations) overlaid on WordNet
Why Additional Relations? Similarity Via Structure-Mapping

Networks of relations are placed in an isomorphic mapping

Sapper: Veale and Keane (1997)
Ad-Hoc Connectivity in Lexical Ontologies: Too Ambitious?

A person who makes maps

Genetics The process of locating genes on a chromosome

Relies on too many felicitous accidents of representation in PWN
Palimpsest is a relational overlay for WordNet. Palimpsest relates word senses in WordNet with simple two-place predicates. Each relation is a Frame-Slot-Filler triple. All Palimpsest relations assume a two-place format (e.g., making a semantic net). The format of Palimpsest relations involves frames and slots, where frame and filler denote PWN senses, and slot represents a relation (active verb). The rationale is "Priming the KR pump": A hand-crafted "skeleton" of basic relationships to be extended automatically.
Rationale: Choose relations to suit the ontological sub-domain

All Palimpsest relations assume a two-place form: \[ R(\text{Subject}, \text{Object}) \]

wears, consumes, enjoys, uses, wields, shoots, throws, performs, practices, executes, participates, lacks_part, lacks_resource, evades, avoid_practice, hinders, nurtures, nurtures_belief, nurtures_disbelief, knows, believes, specializes, comprises, occupies, member, key_figure, key_part, added_part, works_in, manages controls, regulates, inhibits, drives, rides, destroys, kills, breaks, cuts, removes, increases, decreases, affects_part, exhibits, affects_ability, symptom, communicates, carries, sends, transmits, spreads, writes, buys, sells, provides, earns, receives, searches, seeks, desires, made_from, covering, covers, operates

Relations can be hierarchically specialized: e.g., rides/drives(Knight, horse)
An Example: \{Knight\} qua \{male_aristocrat\} in PWN

- **Knight**
  - IS-A: male_aristocrat
  - Rides/drives: horse
  - Wears/uses: armour
  - Wields/uses: sword
  - Exhibits: nobility
  - Practices: chivalry
  - Commands: squire

- **Surgeon**
  - IS-A: doctor
  - Works_in: hospital
  - Wears/uses: gown
  - Wields/uses: scalpel
  - Exhibits: accuracy
  - Practices: surgery
  - Cuts/transforms: flesh

WordNet headword as frame-name

Slot Name as relation

Both typically sharp
Frame Representations as Relation-Object Tuple Sets

**knight**

- Rides/drives: horse
- Wears/uses: armour
- Wields/uses: sword
- Exhibits: nobility
- Practices: chivalry
- Commands: squire

**surgeon**

- Commands: nurse
- Wears/uses: gown
- Wields/uses: scalpel
- Exhibits: accuracy
- Cuts/transforms: flesh
- Practices: surgery

WordNet headword as set-name
## Current Scale of Palimpsest

<table>
<thead>
<tr>
<th>Details</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palimpsest is a relational overlay for WordNet</td>
<td># Concepts</td>
</tr>
<tr>
<td>At present, Palimpsest associates relational triples with</td>
<td>35,514 word senses</td>
</tr>
<tr>
<td>Each relation is a Frame-Slot-Filler triple</td>
<td># Relations</td>
</tr>
<tr>
<td>Palimpsest currently overlays</td>
<td>89,977 relational triples on PWN</td>
</tr>
<tr>
<td>Triple = Subject-sense + Tuple of Relation + Object</td>
<td># Rel-Obj Tuples</td>
</tr>
<tr>
<td>Palimpsest contains 45,415 unique relation-object tuples, avg. 2 per subject</td>
<td></td>
</tr>
<tr>
<td>Sharing, Mapping and “Stretching” Tuples</td>
<td># Matches</td>
</tr>
<tr>
<td>We use Wikipedia associations, + Web-derived associations, to align tuples</td>
<td></td>
</tr>
</tbody>
</table>
Sharing Identical Tuples: Similes

**knight**
- Wears/uses: *armour*
- Exhibits: *nobility*
- Practices: *chivalry*
- Commands: *squire*

**bullfighter**
- Works_in: *bullring*
- Wears/uses: *cape*
- Wields/uses: *sword*
- Rides/drives: *horse*
- Kills/Destroys: *bull*
- Practices: *bull_fighting*
Determining Salience from Wikipedia: A Rich Topological Resource

- Wikipedia comprises a large (growing) collection of articles for headwords. Articles are cross-referenced, yielding an association network of headwords.

- A reverse-Google intuition applies to Wikipedia. The most referenced terms are NOT the most authoritative, but least salient

- References are not weighted, but strong associations can be recognized. Reciprocated cross-linking between headwords A and B indicates high salience.

- Reciprocated and un-Recipricoted Wiki Links in PWN

  PWN 1.6/Wiki-aligned headwords avg. 11 unrecip. links each, 4 recip. Links each

# Wikipedia links
Wikipedia Connectivity indicates Salience of Tuples

** knight **
- Rides/drives: horse
- Wears/uses: armour
- Wields/uses: sword
- Exhibits: nobility
- Practices: chivalry
- Commands: squire

** surgeon **
- Commands: nurse
- Wears/uses: gown
- Wields/uses: scalpel
- Exhibits: accuracy
- Cuts/transforms: flesh
- Practices: surgery
### Graded Levels of Salience

<table>
<thead>
<tr>
<th>Level</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Highest</strong></td>
<td>Explicitly Marked in a Palimpsest relation</td>
</tr>
<tr>
<td>e.g., practices(knight, chivalry)<em>, rides/drives(knight, horse)</em></td>
<td></td>
</tr>
<tr>
<td><strong>High</strong></td>
<td>Reciprocally linked headwords in Wikipedia</td>
</tr>
<tr>
<td>e.g., knight—wiki→armour ∧ armour—wiki→knight</td>
<td></td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
<td>Any unmarked Palimpsest relation</td>
</tr>
<tr>
<td>e.g., occupies(knight, castle), wears/uses(astronaut, helmet)</td>
<td></td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>Singly (not reciprocally) linked headwords in Wikipedia</td>
</tr>
<tr>
<td>e.g., knight—wiki→Germany, chivalry—wiki→horse</td>
<td></td>
</tr>
</tbody>
</table>
Recognizing Similarity with Palimpsest Tuple Representations

**knight**
- Rides/drives: horse
- Wears/uses: armour
- Wields/uses: sword
- Exhibits: nobility
- Commands: squire
- Practices: chivalry

**Samurai**
- Serves: Shogun
- Wields/uses: sword
- Exhibits: nobility
- Practices: bushido

Wiki-reciprocated linkages
Additional Tuples: Web-Derived Attribute-Value Pairings

**peacock**
- **Has_feather:** brilliant
- **Has_plumage:** extravagant
- **Has_strut:** proud
- **Has_tail:** elegant
- **Has_display:** colorful
- **Has_manner:** stately

**lion**
- **Has_gait:** majestic
- **Has_strength:** magnificent
- **Has_soul:** noble
- **Has_eyes:** fierce
- **Has_teeth:** ferocious
- **Has_roar:** threatening
Mining the Web for Diagnostic Attribute-Value Pairings

the ADJ NOUN of a NOUN

as ADJ as a NOUN

the **proud** strut of a peacock

as **proud** as a peacock

the brave heart of a lion

as brave as a lion

the **proud** owner of a peacock

as **proud** as a peacock
Acquire a matrix of “property reinforcement” tendencies from web

<table>
<thead>
<tr>
<th></th>
<th>hot</th>
<th>spicy</th>
<th>humid</th>
<th>fiery</th>
<th>dry</th>
<th>sultry</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>---</td>
<td>35</td>
<td>39</td>
<td>6</td>
<td>34</td>
<td>11</td>
<td>...</td>
</tr>
<tr>
<td>spicy</td>
<td>75</td>
<td>---</td>
<td>0</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>humid</td>
<td>18</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>fiery</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>dry</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>sultry</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>---</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Use the Google query “as * and * as” to acquire associations.
• **Metaphors intuitively seem more creative than similes**
  
  Creativity

  Similes reflect similarity, while Metaphors create/establish similarity

• **Metaphors are essentially Asymmetric**
  
  Asymmetry

  Similes are reversible, but Metaphors change in meaning when reversed

• **Metaphors highlight low-salience features of a topic**
  
  Salience Imbalance

  High-salience features in the vehicle serve to emphasise the same in the topic

• **Metaphor thus performs figure-ground reversal**
  
  Figure Marking

  We thus require a KR with marked levels of salience (i.e., figure vs. ground)
Modelling Simile

The Palimpsest representation of a concept / sense is a Tuple-Set

\[ \text{tuples}(x) \equiv \{<R, o> | R(x, o)\} \quad \text{e.g., tuples(knight) = \{<wields:sword>, ...\}} \]

The basis of literal similarity between two concepts

\[ \text{common}(x, y) \equiv \text{tuples}(x) \cap \text{tuples}(y) \equiv \{<R:o> | R(x, o) \land R(y, o)\} \]

Similes are Symmetrical Comparisons

\[ \text{common}(x, y) \equiv \text{common}(y, x) \]

Similarity is a function of the scope of literal sharing

\[ \text{similarity}(x, y) \propto |\text{common}(x, y)| \propto |\{<R:o> | R(x, o) \land R(y, o)\}| \]
Different Concepts may contain Wiki-Related Tuples

A linguist is like a mathematician

**linguist**
- Specializes_in: *linguistics*
- Specializes_in: *grammar*
- Specializes_in: *language*
- Specializes_in: *syntax*

**mathematician**
- Makes/creates: *equation*
- Makes/creates: *proof*
- Specializes_in: *mathematics*
- Practices: *proof*
- Specializes_in: *algebra*
Different Concepts may contain Wiki-Related Tuples

A bachelor is like a monk

monk

- Works_in: monastery
- performs: prayer
- Avoids_practice: sex
- Practices: celibacy
- Practices: contemplation
- Wears: robe

bachelor

- Lacks_resource: wife
- Avoids_practice: marriage
Modelling Metaphor

The Tacit Tuple-Set of a concept

\[ tacit(c) = \{ <R, o> \mid \neg R(c, o) \land (c \xrightarrow{\text{wiki}} o \land \exists x R(x, o)) \} \]

Metaphoric Mapping

\[ \text{mapped}(\text{vehicle}, \text{tenor}) = (\text{tuples}(\text{tenor}) \cup \text{tacit}(\text{tenor})) \cap \text{tuples}(\text{vehicle}) \]

Metaphoric Projection

\[ \text{projected}(\text{vehicle}, \text{tenor}) = \text{tacit}(\text{tenor}) \cap \text{tuples}(\text{vehicle}) \]

Metaphoric Foregrounding

\[ \text{highlights}(v, t) = \{ <R, o> \in \text{mapped}(v, t) \mid \text{salience}(v, o) > \text{salience}(t, o) \} \]
Analogical Mapping of Tuple Chains: Squaring Rule

Sapper: Veale and Keane (1997)

whaler

Propels/fires: harpoon

Top_part: barb

isa (PWN): projectile

archer

Propels/fires: arrow

Top_part: arrowhead
Analogical Mapping of Tuple Chains: Squaring Rule

archer

Operates/uses: bow

Isa: weapon

propels: arrow

gunner

Operates/uses: gun

propels: bullet
Modelling Analogy

The Extended Tuple-Set of a concept

\[ \text{extended}(c) = \{ <R, e> | R(c, o) \land o \xrightarrow{\text{map}} e \} \]

Object Mapping

\[ o \xrightarrow{\text{map}} e \ \text{IF:} \]

\[ (o \xrightarrow{\text{wiki}} e) \lor (o \xrightarrow{\text{isa}} x \land e \xrightarrow{\text{isa}} x) \lor (R(o, x) \land R(e, x)) \]

Analogical Projection

\[ \text{mapped}(\text{vehicle}, \text{tenor}) = \text{tuples}(\text{tenor}) \cap \text{extended}(\text{vehicle}) \]

Asymmetry

\[ \text{mapped}(x, y) \neq \text{mapped}(y, x) \]
Empirical Considerations: Tuple Re-Use

66% of Palimpsest-annotated concepts can be placed in at least one simile

% of concepts

# of similes per concept
91% of Palimpsest-annotated concepts can be placed in at least one analogy.
Empirical Considerations: Simile Potential

26% of all tuples (11,852 of 45,415) are shared by two or more concepts.

One can generate an average of three similes per tuple.

% of similes

# of shared tuples per simile
Empirical Considerations: Tuple Extendability

% of metaphors

# of mappings per analogy / metaphor

There is an average of eight possible mappings per tuple.

81% of all tuples (i.e., 37,183 of 45,415) can be mapped in one or more metaphors or analogies.
Empirical Considerations: Semantic Distance in Comparisons

Average link-distance between simile nouns = 2 IS-A links

Average link-distance between analogies = 5 IS-A links
Conclusions

- Processes that stretch language stretch onto resources
  - Heterogeneity
  No one resource is currently rich enough to accommodate metaphor

- Hand-crafted knowledge structures can anchor Onto.
  - Palimpsest
  Easier to encode key facts than to cajole them out of limited resources like PWN

- Wikipedia is topical textured dynamic Knowledge source
  - Rich Connectivity
  Provides the “grease” needed to extend and ruggedize a brittle hand-coded KR

- Can be further extended via KR acquisition from web
  - Tomorrow!!!
  Highly diagnostic features/relations can be acquired via linguistic patterns
Syntagmatics & Ontological Enrichment

Tony Veale, Yanfen Hao
School of Computer Science,
Tony.Veale@UCD.ie

UCD Creative Language Systems Group
WordNet: A “lightweight” Ontology for Lexical Semantics

Synsets denote word-senses

But do synsets describe conceptual categories?
WordNet: A “lightweight” System of Non-Categories

Categorization requires each noun-sense have a “radial” category-membership “function”
Problem: Language exploits a Flexible Category Representation

WordNet is a “lightweight ontological” approach to lexical semantics:

But WN lacks a “constructive” semantic representation; others fill this gap:

- **Aristotelian Taxonomies, Description Logics**  
  WordNet, CYC, SUMO
  More Explicit, **BUT** tries to draw sharp lines between overlapping categories

- **Explicit Semantics / First-Order Logic**  
  SUMO, HowNet, CYC
  Supports inference & theorem proving, **BUT** highly selective and often sparse

- **“Firthian” Corpus-Based Approaches**  
  Firth, Sinclair, Hanks
  Ecological sensitivity to word-usage, **BUT** lacks definitive ontological structure
The “Distributional Profile” of a WN term: Syntagmatic Norms

- **Noun used as the subject or object of an active verb** (Agent Noun Verb)
  
  E.g., a virus *infects*, a robot *obeys*, an opera is *composed*, etc.

- **Noun modified by a given adjective** (Attr Noun Adj)
  
  E.g., insults are *hurtful*, clichés are *tired*, priests are *religious*, etc.

- **Noun used in a “Group of X” construction** (Group Noun Noun)
  
  E.g., an *army* of soldiers, a *conclave* of bishops, a *posse* of rappers, etc.

- **Noun used in a PP-phrase with a given prep. head** (Attach Prep Noun)
  
  E.g., *against* an adversary, *via* an intermediary, *along* a channel, etc.
Wikipedia as a Distributional Context: “Virus” and “Infect”

Category dictates behaviour / range of verbs supported

Verb dictates membership in category Virus

On the web: afflatus.ucd.ie (current projects / Lex-Ecologist)
Wikipedia Corpus: Consider the syntagm “Army of X”

- Wikipedia view
  - volunteer = soldier

- from Star Wars?
  - clone = soldier
  - A pleonasm? pretty obvious

- mercenary(238), clone(132), soldier(122), volunteer(72),
  monster(70), robot(63), minion(60), warrior(60), frog(58),
  knight(50), slave(48), demon(46), clansman(46), monkey(46),
  crusader(44), gladiator(38), ant(37),
  lawyer(32), contributor(28), mutant(27), ...

- Microsoft?
  - lawyer = soldier

- open-source view
  - contributor = soldier

- conventional metaphor
  - soldier ant

- Sci-fi bias
  - robot = soldier
A Functional Approach to Lexicography (an example)

(define fundamentalist
  (* (max (%isa arg0 living_thing)
         (%isa arg0 group))
     (min (max (attr arg0 violent)
              (attr arg0 conservative))
         (max (attr arg0 political)
              (attr arg0 religious)))))

One function
Per category
WordNet-based
Membership score

Fuzzy-logic
AND
Adj : Noun
Co-occurrence

OR

Fuzzy-logic

Adj : Noun
Co-occurrence

One function
Per category
WordNet-based
Membership score

Fuzzy-logic
AND
Adj : Noun
Co-occurrence

OR

Fuzzy-logic

Adj : Noun
Co-occurrence
Conceptual Profiling: What is Base and What is Profiled?

Espresso

black
gourmet

concentrated

satisfying

hot
creamy

tasty

trendy
cold

rich

tender

small

intense

strong

How to find these profiled features? How do we exploit them robustly?
Word Senses as Radial Categories \textbf{(our target form)}

\texttt{define espresso}

\begin{align*}
\*(\text{max} \ (\%isa \ arg0 \ content)) \\
\ (%isa \ arg0 \ liquid))
\end{align*}

\begin{align*}
\text{min} \ (\text{max} \ (\text{attr} \ arg0 \ black)) \\
\ (\text{attr} \ arg0 \ dark))
\end{align*}

\begin{align*}
\text{max} \ (\text{attr} \ arg0 \ strong) \\
\ (\text{attr} \ arg0 \ intense) \\
\ (\text{attr} \ arg0 \ concentrated))
\end{align*}
A Functional Approach to Image Schematic Categories

\[
\text{(define path} \\
\quad \text{(combine} \ (* \ 0.2 \ (%\text{isa arg0 path)})) \\
\quad \text{(min} \ \text{agent arg0 connect}) \\
\quad \text{(max} \ \text{attach along arg0}) \\
\quad \text{(attach via arg0)}))
\]

Fuzzy-logic \text{ AND}

Subject : verb collocation

Cut \text{ operator} (\text{> 0 required})

Rewards \text{ Diversity}
Generating Category Definitions by Parsing WordNet glosses

Espresso: “strong black coffee brewed by forcing stream through …”

(define espresso
  (* (\%sim arg0 espresso 3)
      (combine
       (attr arg0 strong)
       (attr arg0 black)
       !
       (* 0.3 (\%isa arg0 coffee))
       *
     )
   )
)

Off-the-shelf WN-based similarity metric

Over 33,000 glosses (for over 61,000 unique word senses) can be shallow parsed to yield a basic category function of this form.

Question: Are the most diagnostic properties of each concept extracted from each gloss?
Automating Category Definition via WordNet glosses II

Lexico-syntactic elements may need to be transformed to fit query language

```
(define white_knight.0
  (* (%sim arg0 white_knight.0 3)
   (combine
     (attr arg0 friendly)
     (performs arg0 acquire)
     !
     (* 0.3 (%isa arg0 company))
   )
  )
)
```

White_Knight: “a company that is a friendly acquirer”
Automating Category Definition via WordNet glosses III

Question: Do the most “useful” concepts have glosses conducive to parsing?

(define typewriter
  (* (%sim arg0 typewriter)
    (combine
      (attr arg0 hand-operated)
      (performs arg0 print)
      !
      (* 0.3(%isa arg0 machine))
      *
    )
  )
)

Typewriter: “a hand-operated machine for printing written messages …”
Automating Category Defns via WordNet Morpho-Semantic Links

WordNet 2.1 contains explicit links among cross-category morphological forms

(define fraternity.0
  (* (\%sim arg0 fraternity.0 3)
    (combine
      (attr arg0 fraternal)
      (attr arg0 brotherly)
      *
    )
  )
)

{fraternity}  \leftrightarrow  {brotherly, fraternal}  \leftrightarrow  {athlete, …}  \leftrightarrow  {athletic, …}
Evaluation: Scale (part I)

• **Hand-Crafted Category-Membership Functions**
  Currently 750 categories (e.g., robot, soldier, army, container, projectile, …)

• **Morpho-Semantic Connections (e.g., athletic/athlete; fraternal/brother)**
  ~ 5000 categories (limited by fixed # of adj : noun mappings)

• **Gloss-derived Category-Membership Functions**
  > 33,000 unique glosses; > 61,000 unique word senses

• **What next?** Attested Similes as a source of most salient Noun properties
  Google: “as * as a NOUN” (e.g., “as stealthy as a ninja”, “as stiff as a corpse”)
The Career of Metaphor: Bowdle and Gentner (2005)

Novel / Original

Metaphors as Comparisons (feature-matching & transfer)

Metaphors as Categorizations (category membership)

Metaphors as Word Senses (e.g., lexicalized in WordNet)

Conventional
The Career of Simile (?)

Explicit Similes
- As hot as an oven
- As flat as a pancake

Non-Explicit Similes
- ... like an oven
- ... like a pancake

Metaophoric Comparisons
- This room is an oven ...
- ... covered with pancake makeup

Novel / Original

Accepted

Conventional
Using Similes to Characterize WordNet Senses / Categories

• Similes / Comparisons reveal the most diagnostic features of a concept
  E.g., “as hot as the sun”, “as dry as sand”, “as wobbly as jelly”, “as sweet as pie”

• The most frequent similes characterize the most pivotal concepts / senses
  E.g., animal concepts (“lion”, “rat”, etc.) are frequently used in comparisons

• Unlike metaphors, similes have a standard, recognizable syntactic frame
  “as barren as a desert”, “as delicate as a surgeon”, “as stiff as a corpse”

• Detailed Knowledge-Representations can be gathered for individual concepts
  Example: surgeon = {delicate, sensitive, skilled, clinical, professional, …}
Web Similes: a cognitive landscape of conceptual landmarks

Many common comparisons are produced / understood w.r.t. these landmarks
Sampling Comparisons/Similes from the WWW

Query-pattern #1: “as ADJ as a/an *” for all antonymous adjectives in WN

Query-pattern #2: “as * as a/an NOUN” for all nouns gathered with query #1

- 200 sampled snippets per query, to give 74,704 apparent simile instances
- 42,618 unique simile types, linking 3769 adjectives to 9287 unique nouns

- Major Issues: Frame Leakage, Implicit/Local Context, Irony
  “as pointed as a question”, “as hairy as a bowling-ball”, “as sober as a Kennedy”

- Clustering of nouns and adjectives supports WN-based sense assignment
  12,259 verified/WSD simile types, of 2124 adjectives to 3778 noun senses

View on the Web: http://afflatus.ucd.ie/sardonicus/tree.jsp
Most common noun vehicles in bona-fide non-ironic similes.

Note familiarity / basic-level of the vehicles.

[rock (55), statue (46), cat (45), puppy (43), tiger (43), baby (42),
diamond (41), child (40), snake (39), lion (37), hurricane (37), lamb (36),
kitten (35), cloud (34), corpse (34), mountain (33), wolf (33), dream (32),
snowflake (31), tornado (31), dog (30), earthquake (28), butterfly (26),
dove (26), lover (25), nun (25), mother (25), teddy_bear (25), bear (25),
eagle (24), tree (24), panther (24), virus (24), pig (24), soap_opera (23),
mirror (23), cow (23), bull (23), pearl (23), queen (23), clock (22),
dolphin (22), rabbit (22), serpent (21), peach (21), bride (21),
flower (21), king (21), mule (21), judge (21), peacock (21), fox (21),
storm (21), tomb (21), leopard (21), mouse (21), hummingbird (21),
glacier (20), thunderstorm (20), sheep (20), hawk (20), computer (20),
gazelle (20), monk (19), brick (19), tank (19), bullet (19), surgeon (19),
supermodel (18), clam (18), roller_coaster (18), crocodile (18),
freight_train (18), oak (18), worm (18), horse (18), bird (17),
mannequin (17), feather (17), bulldog (17), rocket (17), prince (17),
warrior (17), infant (17), sparrow (17), photograph (17), robot (17),
encyclopedia (17), shark (17), sponge (17), oak_tree (17), lark (16),
swan (16), toddler (16), bunny (16), ballerina (16), whale (16),
monkey (16), painting (15), soap_bubble (15), mirage (15),
dictionary (15), sunrise (15), babe (15), schoolgirl (15), poet (15),
graveyard (15), cobra (15), laser (15), bomb (15), apple (14), mummy (14),
lioness (14), politician (14), rattlesnake (14), pup (14), elephant (14),]
Most common adjectival predicates for bona-fide non-ironic similes.
Ironic Comparisons/Similes from the WWW

Some Examples:

As \{welcome, painless, appealing, pleasant, exciting, entertaining\} as a root-canal
As subtle as a \{sledgehammer, freight_train, anvil, axe, rhino, toilet_seat, ...\}
As hefty as a \{laptop, croissant\}
As blind as a \{referee, hawk\}
As \{muscular, epicurean, smart, straight, sturdy, weighty, ...\} as a paper_clip
As rare as a \{ham_sandwich, toaster, traffic_jam, monsoon, garbage_pickup\}
As \{bulletproof, scary, subversive\} as a sponge_cake
As private as a \{park_bench, town_hall, shopping_mall\}

View on the Web: \texttt{http://afflatus.ucd.ie/sardonicus/tree.jsp}

2796 unique adj:noun ironic simile types.
936 adjectives to 1417 nouns.
13% of all annotated simile instances. 18% of unique simile types.
Most common noun vehicles in ironic similes.

Note increased level of specificity and lower density of reuse.
Most common adjectival predicates for ironic similes.

Note prevalence of positive adjectives.
Generating Category Functions from Simile-derived features

(define ninja.0
  (* (%isa arg0 person.0)
    (* (%sim arg0 ninja.0)
      (combine
       (attr arg0 agile)
       (attr arg0 deadly)
       (attr arg0 stealthy)
       (attr arg0 graceful)
       (attr arg0 silent))))
)
)

Question: “Where does the hard constraint of “person-hood” come from?
What Other Concepts Can be a Category be Stretched to Include?

Metaphor is a process of “Category Inclusion” (Glucksberg, 2001)

Queries of form “CAT-like person” reveal extent that CAT can include people

Generalizing across all 3778 WN senses, we see the following inclusion patterns

<table>
<thead>
<tr>
<th>X like Y</th>
<th>Person</th>
<th>Animal</th>
<th>Substance</th>
<th>Tool</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>66%</td>
<td>5%</td>
<td>3%</td>
<td>4%</td>
<td>9%</td>
</tr>
<tr>
<td>Animal</td>
<td>36%</td>
<td>27%</td>
<td>4%</td>
<td>5%</td>
<td>15%</td>
</tr>
<tr>
<td>Substance</td>
<td>14%</td>
<td>3%</td>
<td>37%</td>
<td>5%</td>
<td>32%</td>
</tr>
<tr>
<td>Tool</td>
<td>8%</td>
<td>3%</td>
<td>7%</td>
<td>22%</td>
<td>34%</td>
</tr>
<tr>
<td>Structure</td>
<td>4%</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
<td>43%</td>
</tr>
</tbody>
</table>

E.g., “snake-like structure” (125 hits), “snake-like animal” (122 hits) ...
### Overall Word Appreciation: Whissell’s Dictionary of Affect

<table>
<thead>
<tr>
<th>Word</th>
<th>Pleasantness</th>
<th>Activation</th>
<th>Imagery</th>
</tr>
</thead>
<tbody>
<tr>
<td>ugly</td>
<td>1.0000</td>
<td>1.8333</td>
<td>2.2</td>
</tr>
<tr>
<td>ultimate</td>
<td>2.1667</td>
<td>1.8000</td>
<td>1.2</td>
</tr>
<tr>
<td>umbilical</td>
<td>1.7500</td>
<td>1.6667</td>
<td>2.6</td>
</tr>
<tr>
<td>unable</td>
<td>1.0000</td>
<td>1.6000</td>
<td>1.4</td>
</tr>
<tr>
<td>unanswered</td>
<td>1.2857</td>
<td>1.1667</td>
<td>1.4</td>
</tr>
<tr>
<td>unaware</td>
<td>1.1429</td>
<td>1.3333</td>
<td>1.6</td>
</tr>
<tr>
<td>unburdened</td>
<td>2.3750</td>
<td>1.4286</td>
<td>1.8</td>
</tr>
<tr>
<td>uncertain</td>
<td>1.3333</td>
<td>1.6000</td>
<td>1.4</td>
</tr>
<tr>
<td>uncertainty</td>
<td>1.2857</td>
<td>1.6667</td>
<td>1.4</td>
</tr>
<tr>
<td>uncle</td>
<td>2.0000</td>
<td>1.8000</td>
<td>3.0</td>
</tr>
<tr>
<td>uncombed</td>
<td>1.3333</td>
<td>1.6250</td>
<td>2.8</td>
</tr>
<tr>
<td>uncomfortable</td>
<td>1.0000</td>
<td>1.5714</td>
<td>1.8</td>
</tr>
<tr>
<td>unconscious</td>
<td>1.3750</td>
<td>1.0000</td>
<td>2.2</td>
</tr>
<tr>
<td>unconsciously</td>
<td>1.5000</td>
<td>1.2000</td>
<td>1.6</td>
</tr>
</tbody>
</table>

~ 8000 words (all synonyms) with numeric dimensions based on volunteer ratings

Mean pleasantness = 1.85 (standard dev. = 0.36)  
3.0 = best, 1.0 = worst

Realistic Assumption: Pleasantness rating is based on an overall understanding of a word/concept
Estimating Affect: Pleasantness depends on Salient Features

Snake = [slippery(8), supple(6), cunning(5), dangerous(3), poisonous(1) ...]

⇒

\[affect_{est}(\text{snake}) \approx 8 \times affect(\text{slippery}) + 6 \times affect(\text{supple}) + 5 \times affect(\text{cunning}) + \ldots\]

\[8 + 6 + 5 + 3 + 1 + \ldots\]

If most diagnostic features have been acquired:

We expect a strong correlation between affect_{est} and affect_{Whissell}
Empirical Evaluation: Predicting the Affect of a Concept

Pearson two-tailed correlation coefficient: estimated affect with Whissell affect

Correlation

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3

All simile features
+0.35

Ironic only
-0.243

Non-ironic features only
+0.514

PWN gloss features
+0.278

All wiki features
+0.15
Empirical Evaluation: Determining Descriptive Sufficiency

How well do web-gathered salient properties describe a given concept?

Expectation: should predict affective ratings of nouns in Whissell's dictionary ('89)

- All web features yield a predictive accuracy of 0.35
  - Straight similes: correlation of 0.514
  - Ironic alone: correlation of -0.243

- All adjectival modifiers for the noun concept in a corpus
  - Wikipedia Corpus
  - Noisy, biased, extrinsic features? Yields a correlation of just 0.15

- All adjectival modifiers found from WordNet glosses
  - Gloss features
  - Surprisingly unpredictive: yields a correlation of just 0.278
Almuhareb & Poesio (2004): Clustering Concepts by Modifiers/Attributes

<table>
<thead>
<tr>
<th>Class</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, mouse, oyster, puppy, rat, sheep, tiger, turtle, zebra</td>
</tr>
<tr>
<td>Building</td>
<td>abattoir, center, clubhouse, dormitory, greenhouse, hall, hospital, hotel, house, inn, library, nursery, restaurant, school, skyscraper, tavern, theater, villa, whorehouse</td>
</tr>
<tr>
<td>Cloth</td>
<td>pants, blouse, coat, costume, gloves, hat, jacket, jeans, necklace, pajamas, robe, scarf, shirt, suit, trousers, uniform</td>
</tr>
<tr>
<td>Creator</td>
<td>architect, artist, builder, constructor, craftsman, designer, developer, farmer, inventor, maker, manufacturer, musician, originator, painter, photographer, producer, tailor</td>
</tr>
<tr>
<td>Disease</td>
<td>acne, anthrax, arthritis, asthma, cancer, cholera, cirrhosis, diabetes, eczema, flu, glaucoma, hepatitis, leukemia, malnutrition, meningitis, plague, rheumatism, smallpox</td>
</tr>
<tr>
<td>Feeling</td>
<td>anger, desire, fear, happiness, joy, love, pain, passion, pleasure, sadness, sensitivity, shame, wonder</td>
</tr>
<tr>
<td>Fruit</td>
<td>apple, banana, berry, cherry, grape, kiwi, lemon, mango, melon, olive, orange, peach, pear, pineapple, strawberry, watermelon</td>
</tr>
<tr>
<td>Furniture</td>
<td>bed, bookcase, cabinet, chair, couch, cradle, desk, dresser, lamp, lounge, seat, sofa, table, wardrobe</td>
</tr>
<tr>
<td>Body Part</td>
<td>ankle, arm, ear, eye, face, finger, foot, hand, head, leg, nose, shoulder, toe, tongue, tooth, wrist</td>
</tr>
<tr>
<td>Family Relation</td>
<td>boy, child, cousin, daughter, father, girl, grandchild, grandfather, grandmother, husband, kid, mother, offspring, sibling, son, wife</td>
</tr>
<tr>
<td>Time</td>
<td>century, decade, era, evening, fall, hour, month, morning, night, overtime, quarter, season, semester, spring, summer, week, weekend, winter, year</td>
</tr>
<tr>
<td>Vehicle</td>
<td>aircraft, airplane, automobile, bicycle, boat, car, cruiser, helicopter, motorcycle, pickup, rocket, ship, truck, van</td>
</tr>
</tbody>
</table>

214 concepts from 13 PWN categories
Almuhareb & Poesio (2004): Web-Mining of Concept Modifiers/Attributes

Finds 51,045 modifiers for 214 nouns

Google query to mine adjectival modifiers for a given Concept

Finds 8934 attributes for 214 nouns

e.g., rocket = [fast, powerful, speed, thrust, …]  vector space of 59,979 features
Recall Mining the Web for Diagnostic Attribute-Value Pairings

* the ADJ NOUN of a NOUN
  as ADJ as a NOUN

the proud strut of a peacock
as proud as a peacock

the brave heart of a lion
as brave as a lion

the proud owner of a peacock
as proud as a peacock
### Clustering Results

**13-way clustering: \([I2=9.58e+001] [214 of 214]\), Entropy: 0.133, Purity: 0.902**

| cid | Entpy | Purty | body | crea | dise | fami | vehi | publ | feel | clot | buil | time | anim | frui | furn |
|-----|-------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 0   | 0.000 | 1.000 | 0    | 0    | 18   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 1   | 0.087 | 0.941 | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 16   |
| 2   | 0.106 | 0.923 | 0    | 1    | 0    | 0    | 0    | 0    | 12   | 0    | 0    | 0    | 0    | 0    | 0    |
| 3   | 0.000 | 1.000 | 0    | 13   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 4   | 0.000 | 1.000 | 0    | 16   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 5   | 0.000 | 1.000 | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 6   | 0.321 | 0.750 | 0    | 1    | 0    | 0    | 12   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 17   |
| 7   | 0.160 | 0.895 | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 0    | 1    | 0    | 0    |
| 8   | 0.100 | 0.929 | 0    | 1    | 0    | 13   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 9   | 0.000 | 1.000 | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 10  | 0.155 | 0.864 | 0    | 0    | 0    | 3    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 19   |
| 11  | 0.405 | 0.722 | 0    | 0    | 0    | 1    | 1    | 1    | 0    | 1    | 0    | 1    | 1    | 0    | 0    |
| 12  | 0.286 | 0.789 | 0    | 1    | 0    | 0    | 0    | 15   | 0    | 0    | 2    | 1    | 0    | 0    | 0    |

**Compare 0.855 for Almuhareb & Poesio (2004)**

**Compare V+H:**
- 7183 feat.

**Compare A+P:**
- 59,979 feat.
Visualizing Concept Clusters based on Diagnostic Features

Furniture

Body_part
### Recognizing Irony: Affective Patterns across Comparisons

<table>
<thead>
<tr>
<th></th>
<th>Pos-Adj</th>
<th>Neg-Adj</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-Ironic (bona-fide) similes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Affect Noun</td>
<td>48.1%</td>
<td>16.6%</td>
</tr>
<tr>
<td>Negative Affect Noun</td>
<td>19.9%</td>
<td>15.4%</td>
</tr>
<tr>
<td><strong>Ironic Similes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Affect Noun</td>
<td>30.9%</td>
<td>15.3%</td>
</tr>
<tr>
<td>Negative Affect Noun</td>
<td>43.1%</td>
<td>10.7%</td>
</tr>
</tbody>
</table>

- **Positive:** Affect $\geq 1.72$
- **Negative:** Affect $< 1.72$

2288 similes have an adjective and noun in Whissell’s dictionary.
Domain Incongruence: Understanding Properties in Context

Domain Incongruence: Properties have different meanings in the tenor domain. E.g., “a theory as leaky as a sieve”, “theory as strong as a rock”

- Problem: How to know the preferred choice of adjectives for a tenor
  Noun Description Vector: all adjectives used to modifier a noun in our corpus

- Problem: Understand how that properties reinforce/imply each other
  Property Co-description Vector: all bindings for “as ADJ and * as” on Google

- Problem: Which local properties are implied for a NOUN by a given ADJ?
  Solution: interpret(Adj : Noun) = description(Noun) \(\cap\) co-description(Adj)
Recall **Slippage and Implication matrix** for Adjectives

Acquire a matrix of “property reinforcement” tendencies from web

<table>
<thead>
<tr>
<th></th>
<th>hot</th>
<th>spicy</th>
<th>humid</th>
<th>fiery</th>
<th>dry</th>
<th>sultry</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>---</td>
<td>35</td>
<td>39</td>
<td>6</td>
<td>34</td>
<td>11</td>
<td>...</td>
</tr>
<tr>
<td>spicy</td>
<td>75</td>
<td>---</td>
<td>0</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>humid</td>
<td>18</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>fiery</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>dry</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>sultry</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>---</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Uses the Google query

“as * and * as”

to acquire associations
Example: Interpret the description “a flimsy theory”

Flimsy theory =

{weak(3), self-serving(3), predictable(3), ridiculous(2), subjective(2), dramatic(2), unconvincing(1), incredible(1), spurious(1), small(50), arbitrary(48), implausible(46), preposterous(46), tenuous(46), silly(44)}
Robust theory =

\{\text{effective}(18), \text{powerful}(13), \text{rigorous}(13), \\
\text{durable}(10), \text{general}(9), \text{comprehensive}(8), \\
\text{accurate}(7), \text{functional}(7), \text{solid}(6), \\
\text{energetic}(5), \text{sophisticated}(5), \text{strong}(4), \\
\text{competitive}(4), \text{integrated}(4), \text{dynamic}(4), ...\}
Example: Interpret the description “a shaky theory”

Shaky\_theory =

\{weak(2), inadequate(2), tenuous(2), weird(1),
 controverisal(1), old(1), subjective(1),
 problematic(1), ridiculous(1), ... \}

Watertight\_theory =

\{robust(1), effective(1), accurate(1),
 solid(1), objective(1), comprehensive(1)\}
Conclusions

• The knowledge needed to reorganize WN is not to be found in WN itself. WN glosses do not indicate the most diagnostic/salient properties of a concept.

• Yet, lightweight ontologies like WordNet can be converted automatically. Formulating diagnostic properties as corpus-sensitive queries contextualizes WN.

• Similes provide best clues to the salient properties of figurative vehicles. A large case-base of “natural” comparisons is easily acquired from the web.

• Metaphor/Simile Processing On-Line

  Generate metaphors for arbitrary target concepts that highlight given features.

  [afflatus.ucd.ie/aristotle]
Cross-Resource Semantic Insights via Orthographic Decomposition

Tony Veale
School of Computer Science,
UCD (NUI-D)
Tony.Veale@UCD.ie
Morphological Structure and Meaning

**Morphemes are the smallest units of linguistic meaning**

A morphemic analysis of words reveals insights into semantic composition

- **Multi-morphemic words reveal a language's history**
  
  E.g., “Reimburse” = Re + Im + Burse = BACK + INTO + PURSE

- **Shared Morphemes indicate shared Meaning**
  
  E.g., “Miracle” / “Mirage” / “Admire” all derive from Mirare (to “wonder”)

- **Latin / Greek derivations are most decomposable**
  
  But consider Anglo-German: Zeitgeist = “Zeit” + “Geist” = TIMES + GHOST
Typical Taxonomic Structure in a Lexical Ontology

![Ontology Diagram](attachment:image.png)
Mathematician | 数学家 = Mathematics | 数学 + Human | 人
Unlocking Chinese Orthography: Latent Semantic Cues in HowNet
Unlocking Chinese Orthography: Latent Semantic Cues in HowNet

Exercise - Human

Baseball - Human

Ballplayer

Gymnast

Athletics - Human

Mountaineering - Human

Mountaineer

Pole Vaulting - Human

Pole Vaulter

Fencing - Human

Swordsman
# Uses of Orthographic Decomposition

<table>
<thead>
<tr>
<th>Uses</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equivalence</strong></td>
<td>$A + B \Rightarrow A = B$</td>
</tr>
<tr>
<td>E.g., (voice) 声音 = (voice) 声 + (sound)音 $\Rightarrow$ (voice)声 = (sound)音</td>
<td></td>
</tr>
<tr>
<td><strong>Pure Composition</strong></td>
<td>$A + B \Rightarrow AB$</td>
</tr>
<tr>
<td>E.g., (scalpel)手术刀 = (surgery) 手术 + (knife)刀</td>
<td></td>
</tr>
<tr>
<td><strong>Emergence</strong></td>
<td>$A + B \Rightarrow C$</td>
</tr>
<tr>
<td>E.g., (hair or feather) 毛 + (sickness) 病 $\Rightarrow$ (shortcoming)缺点</td>
<td></td>
</tr>
<tr>
<td><strong>Modification</strong></td>
<td>$A + B \Rightarrow A \text{ modifies } B$</td>
</tr>
<tr>
<td>E.g., (espresso) 浓咖啡 $\Rightarrow$ (strong) 浓 + (coffee) 浓 咖</td>
<td></td>
</tr>
</tbody>
</table>
**Forms of Orthographic Decomposition: Nouns**

- **AB(noun) = A(adj) B(noun) ⇒ B has property A**
  \[\text{Adj : Noun}\]
  
  E.g., (samurai) 武士  = (valiant) 武  + (scholar) 士  \(\text{valiant + educated}\)

- **AB(noun) = A(verb) B(noun) ⇒ B performs A**
  \[\text{Verb : Noun}\]
  
  E.g., (knight) 騎士  = (ride) 騎  + (scholar) 士  ⇒ *A knight rides*

- **AB(noun) = A(noun) B(noun) ⇒ B performs A**
  \[\text{Noun : Noun}\]
  
  E.g., (fighter) 战鹰  = (war) 战  + (hawk) 鹰  ⇒ *a fighter is a “war hawk”*

- **ABC(noun) = A(verb) B(noun) C(noun) ⇒ C does A to C**
  \[\text{(Verb:Noun):Noun}\]
  
  E.g., (vampire) 吸血鬼  = (suck) 吸  + (blood) 血  + (ghost) 鬼
HowNet: Property / Modifier Taxonomy

E.g., (samurai)武士 ⇒ Samurai:CourageValue=valiant 武
**Decomposition Unlocks Implicit World Knowledge**

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Decomposition</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown rice</td>
<td>米 is already a phrasal decomposition in English</td>
<td>brown rice : smoothness = rough</td>
</tr>
<tr>
<td>(Blueprint)草案</td>
<td>(rough)草 + (draft)案</td>
<td>blueprint : carefulness = rough</td>
</tr>
<tr>
<td>(estimate)概算</td>
<td>(rough)概 + (count)算</td>
<td>estimate : accuracy = rough</td>
</tr>
<tr>
<td>(robot)机器人</td>
<td>(mechanical)机 + (person)人</td>
<td>robot like a person</td>
</tr>
<tr>
<td>(bone-joint)节</td>
<td>(skeleton)骨 + (knot)节</td>
<td>a joint is like a knot</td>
</tr>
<tr>
<td>(breast)乳房</td>
<td>(milk)乳 + (house)房</td>
<td></td>
</tr>
<tr>
<td>(sky)天宇</td>
<td>(celestial)天 + (house)字</td>
<td>Kenning Riddles</td>
</tr>
</tbody>
</table>
### Aligning HowNet and WordNet: Multi-Strategy Approach

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synonymy</strong></td>
<td>$xyz \rightarrow a \land xyz \rightarrow b \land {a, b} \in \text{WordNet}$</td>
</tr>
<tr>
<td></td>
<td>E.g., $_ \rightarrow \text{“head”} \land _ \rightarrow \text{“chief”} \land {\text{head, chief}} \in \text{WordNet}$</td>
</tr>
<tr>
<td><strong>Polysemy</strong></td>
<td>$xyz \rightarrow a \land xyz \rightarrow b \land \text{gloss}{a, \ldots} = \text{“... b ...”}$</td>
</tr>
<tr>
<td></td>
<td>E.g., $_ \rightarrow \text{“pen”}, \text{“enclosure”} \land \text{gloss}{\text{pen}} = \text{“an enclosure for ...”}$</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>$xyz \rightarrow a \land xyz \rightarrow b \land {a, \ldots} \text{ ISA } {b,\ldots}$</td>
</tr>
<tr>
<td></td>
<td>E.g., $__ \rightarrow \text{“scalpel”}, \text{“surgical knife”} \land {\text{scalpel}} \text{ ISA } {\text{surgical knife}}$</td>
</tr>
<tr>
<td><strong>Compounding</strong></td>
<td>$xyz \rightarrow a \land yz \rightarrow b \land {a, \ldots} \text{ ISA } {b,\ldots}$</td>
</tr>
<tr>
<td></td>
<td>E.g., 语言学家 $\rightarrow \text{“linguist”} \land \text{学家} \rightarrow \text{“scientist”}$</td>
</tr>
</tbody>
</table>
### Aligning Chinese and WordNet: Empirical Results

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Precision (all words)</th>
<th>Recall (all words)</th>
<th>Precision (multi-char)</th>
<th>Recall (multi-char)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy</td>
<td>.79</td>
<td>.12</td>
<td>.96</td>
<td>.20</td>
</tr>
<tr>
<td>Generalization</td>
<td>.73</td>
<td>.09</td>
<td>.88</td>
<td>.15</td>
</tr>
<tr>
<td>Polysemy</td>
<td>.66</td>
<td>.11</td>
<td>.74</td>
<td>.15</td>
</tr>
<tr>
<td>Compounding</td>
<td>.63</td>
<td>.13</td>
<td>.87</td>
<td>.16</td>
</tr>
<tr>
<td>Mod Compound</td>
<td>.61</td>
<td>.11</td>
<td>.62</td>
<td>.41</td>
</tr>
<tr>
<td>Head Compound</td>
<td>.87</td>
<td>.04</td>
<td>.87</td>
<td>.18</td>
</tr>
<tr>
<td>Any two</td>
<td>.74</td>
<td>.17</td>
<td>.89</td>
<td>.24</td>
</tr>
<tr>
<td>Any one</td>
<td>.64</td>
<td>.58</td>
<td>.64</td>
<td>.56</td>
</tr>
</tbody>
</table>

As measured against 5000 hand-aligned HowNet/PWN entries
### Aligning Korean and WordNet: Empirical Results

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Precision (all words)</th>
<th>Recall (all words)</th>
<th>Precision (multi-char)</th>
<th>Recall (multi-char)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy</td>
<td>.90</td>
<td>.28</td>
<td>.90</td>
<td>.30</td>
</tr>
<tr>
<td>Generalization</td>
<td>.91</td>
<td>.23</td>
<td>.91</td>
<td>.23</td>
</tr>
<tr>
<td>Polysemy</td>
<td>.71</td>
<td>.23</td>
<td>.71</td>
<td>.24</td>
</tr>
<tr>
<td>Compounding</td>
<td>.73</td>
<td>.27</td>
<td>.73</td>
<td>.31</td>
</tr>
<tr>
<td>Mod Compound</td>
<td>.69</td>
<td>.18</td>
<td>.70</td>
<td>.21</td>
</tr>
<tr>
<td>Head Compound</td>
<td>.80</td>
<td>.05</td>
<td>.80</td>
<td>.16</td>
</tr>
<tr>
<td>Any two</td>
<td>.83</td>
<td>.36</td>
<td>.83</td>
<td>.37</td>
</tr>
<tr>
<td>Any one</td>
<td>.73</td>
<td>.62</td>
<td>.74</td>
<td>.63</td>
</tr>
</tbody>
</table>

As measured against 12,000 hand-aligned Korterm entries
Conclusions

• **Chinese provides a means of “Creative Decomposition” for lexical concepts**
  
  This crossover knowledge can be used in an English-only app (e.g., a thesaurus)

• **Decomposition yields tired metaphors in Chinese / Fresh views in English**
  
  E.g., tractor = iron + ox  gardener = flower + artisan

• **Corpus-based (as opposed to semantic-based) filters are required**
  
  We want to preserve, rather than filter, “unusual” decompositions

• **Application: The Analogical Thesaurus**  [http://afflatus.ucd.ie](http://afflatus.ucd.ie)
  
  A browsable database of creative synonyms and sim-based term clusters